Citation: Clin Transl Sci (2018) XX, 1–10; doi:10.1111/cts.12602

ARTICLE

Evaluation of Wearable Digital Devices in a Phase I Clinical Trial

Elena S. Izmailova[1,*], Ian L. McLean[1], Gaurav Bhatia[2], Greg Hather[1], Matthew Cantor[2], David Merberg[1], Eric D. Perakslis[1], Christopher Benko[2] and John A. Wagner[1]

We assessed the performance of two US Food and Drug Administration (FDA) 510(k)-cleared wearable digital devices and the operational feasibility of deploying them to augment data collection in a 10-day residential phase I clinical trial. The Phillips Actiwatch Spectrum Pro (Actiwatch) was used to assess mobility and sleep, and the Vitalconnect HealthPatch MD (HealthPatch) was used for monitoring heart rate (HR), respiratory rate (RR), and surface skin temperature (ST). We measured data collection rates, compared device readouts with anticipated readings and conventional in-clinic measures, investigated data limitations, and assessed user acceptability. Six of nine study participants consented; completeness of data collection was adequate (> 90% for four of six subjects). A good correlation was observed between the HealthPatch device derived and in-clinic measures for HR (Pearson $r = 0.71$; $P = 2.2e\text{-}16$) but this was poor for RR ($r = 0.08$; $P = 0.44$) and ST ($r = 0.14$; $P = 0.14$). Manual review of electrocardiogram strips recorded during reported episodes of tachycardia > 180 beats/min showed that these were artefacts. The HealthPatch was judged to be not fit-for-purpose because of artefacts and the need for time-consuming manual review. The Actiwatch device was suitable for monitoring mobility, collecting derived sleep data, and facilitating the interpretation of vital sign data. These results suggest the need for fit-for-purpose evaluation of wearable devices prior to their deployment in drug development studies.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?

☑ Wearable sensors have the potential to collect health-related data remotely, thus enabling acquisition of dense physiological study subject profiles, allowing data collection on an outpatient basis, and thereby reducing the number of clinical study hospital or clinical pharmacology unit (CPU) visits.

WHAT QUESTION DID THIS STUDY ADDRESS?

☑ We evaluated the performance of two FDA 510(k)-cleared devices, HealthPatch MD by Vitalconnect and Actiwatch Spectrum Pro by Phillips, for continuous physiological data collection, compared device readouts with conventional analogous measures and published data, and assessed operational feasibility in a residential phase I clinical trial.

WHAT DOES THIS STUDY ADD TO OUR KNOWLEDGE?

☑ The Actiwatch device was suitable for monitoring mobility, collecting derived sleep data, and providing metadata for interpreting vital sign data. The HealthPatch device was not determined to be "fit-for-purpose" because of the artefacts and the need of extensive, time-consuming manual data review.

HOW MIGHT THIS CHANGE CLINICAL PHARMACOLOGY OR TRANSLATIONAL SCIENCE?

☑ Our study results indicate the need for evaluation of wearable digital device according to fit-for-purpose principle in the context of clinical investigations.

Despite the widespread adoption of consumer digital technologies and their increasing use in healthcare settings, they have yet to find widespread application in industry-sponsored drug development. Some progress has been made, including pilot studies for remotely run clinical trials,[1–3] novel technological solutions to improve medication adherence,[4] and multiple modalities of using digital sensors to create new data streams to improve the collection of health-related data.[5] Despite these efforts, published reports of study results remain limited. Moreover, the results of some studies indicate that digital innovation in health care is more complicated than anticipated[6] and that some technologies do not perform as the researchers had planned.[7] In addition, published results indicate that there is a need for extensive manual review of data and for the investigation of potential device-derived data artefacts, activities that can be time consuming.[8]

[1]Takeda Pharmaceuticals International, Inc., Cambridge, Massachusetts, USA; [2]Koneksa Health, Inc., New York, New York, USA. *Correspondence: Elena S. Izmailova (elena.s.izmailova@gmail.com)

Received: 18 September 2018; accepted: 26 October 2018; published online on XX XX XXXX. doi:10.1111/cts.12602

In early-stage drug development clinical trials, vital sign data, such as heart rate (HR) and respiratory rate (RR), are typically collected manually by clinical personnel or by using electronic devices that record these data at discrete single time points. This is generally done at predefined times before, during, and after administration of the study drug, while a subject is a resident at a clinical pharmacology unit (CPU) or returns for follow-up visits. Additional assessments of vital sign data are performed in response to suspected safety or tolerability issues or if the study drug or a challenge agent is expected to have pharmacological effects on vital signs. High-density vital sign data recorded continuously using wearable digital sensors have the potential (i) to provide more information on study subjects' physiological profiles and greater sensitivity for detecting changes in these parameters, (ii) to allow periods of data collection to be done in the outpatient/home setting rather than as inpatients in the residential CPU setting, and (iii) as an aid in interpretation of adverse events, with an overall view to reducing the time of residential observation during phase I studies and the number of follow-up clinic visits needed.

Wearable digital devices may also have utility for evaluating the impact of a novel medicine on disease activity or outcomes. In many therapeutic indications, the impact of a drug on activities of daily living, including physical activity levels and sleep patterns, is captured routinely as an indication of the potentially clinically relevant benefit of therapy or potential negative side effects. These assessments usually rely on the subject's ability to subsequently recall these events for self-completed questionnaires. This type of data can be subjective, vague, and prone to confounding and bias,[9] aspects that may be improved by continuous real-time collection of activity-related data by digital devices to objectively monitor activities of daily living.

The selection and deployment of appropriate wearable digital devices in the context of drug development presents challenges[5] that are similar to those encountered with the introduction of novel laboratory biomarker tests in the early 2000s. To address these challenges, the "fit-for-purpose" concept was developed by the American Association of Pharmaceutical Scientists (AAPS) Biomarker Workshop[10] and advanced further by the US Food and Drug Administration (FDA)-National Institutes of Health (NIH) Biomarkers Endpoints and other Tools (BEST) working group.[11] Using this framework, a potential biomarker should be evaluated for a predefined purpose in the context that it will be used. We applied this approach to evaluate two wearable digital devices that have 510(k) clearance from the FDA: the Phillips Actiwatch Spectrum Pro (Actiwatch) and the Vitalconnect HealthPatch MD (HealthPatch). We incorporated the testing of these devices as an exploratory component in a 10-day residential phase I study recruiting normal healthy volunteers.

The goal of this substudy was to evaluate whether the HealthPatch and Actiwatch devices were fit-for-purpose to enhance vital sign data collection and to capture physical activity in the context of an industry-sponsored early-phase drug development study. Aspects examined included (i) a comparison with the traditional conventional measures performed at the clinical site, (ii) assessment of a diurnal variation of physiological parameters that were expected to conform to expected temporal patterns, and (iii) understanding data limitations and technical issues. We also assessed the operational aspects of device use, including acceptability for the study subjects and the site personnel.

METHODS

The clinical study was conducted at a US-based single-site residential CPU for a 10-day period. All subjects were healthy volunteers recruited from the CPU's panel; they had no clinically significant acute or chronic medical disorders, were taking no concomitant medications, and had no exposure to other investigational agents in the 30 days preceding the study. The devices were deployed during the CPU confinement period only. Informed consent was obtained separately for the device component of the study, which was optional for any subject consenting to participate in the core part of the study. The study conduct was reviewed and approved by the institutional review board.

For the design of the study and authoring of the study protocol, the vital signs, activity, and sleep data produced by wearable devices were treated as "exploratory," used for device evaluation purpose only, and not linked to primary or secondary study endpoints, which included pharmacokinetic and safety assessments. The data were not available to CPU or sponsor staff during the conduct of the study and were not intended to guide clinical care or other decision making.

The Actiwatch[12] was worn on the wrist using a standard wristwatch-style strap and captured data on motion using an accelerometer, which were used to derive information on activity level and sleep. Activity level is summarized using activity counts, a dimensionless measure of motion that removes the effects of gravity, transportation, and other acceleration not indicative of activity. The HealthPatch[13] was applied to the anterior surface of the left upper precordium using an adhesive strip and captured biometric data: HR, RR, skin temperature (ST), and step count. Both devices were intended to be worn throughout the entire 10-day period of confinement in the CPU. At the end of the study, the site personnel and the study participants were asked to complete a satisfaction questionnaire.

Device data collection

The data collected by the Actiwatch were retrieved by periodically connecting it to a laptop computer running study-specific software, which downloaded the epoch level data from the device to the computer before transfer to the Philips database (**Figure S1**). The HealthPatch device recorded a single-lead anterior chest wall echocardiogram (ECG) voltage every 8 ms, and from the resulting R-R interval an estimate of HR was calculated approximately every 4 seconds, averaging 15 HR estimates within a minute. The data collected by the HealthPatch were streamed from the HealthPatch to a companion iPhone application (Healthwatch, version 2.5.4) on a dedicated iPhone 5 via Bluetooth technology (**Figure S1**).

Device data processing

HealthPatch data were first subjected to a quality control step during which invalid readings were filtered using the manufacturer's proprietary software.

To facilitate estimation of data completeness, gap thresholds were defined. This threshold (T) was set to 5 and 30 seconds for HealthPatch and Actiwatch, respectively. Then, for each device–subject data stream, data were sorted in timestamp order, and the intervals between valid recordings were calculated. If an interval was greater than the gap threshold, it was considered a gap (i.e., missing data). Total noncovered time was calculated by summing the length of all gaps. Percent completeness was defined as 100% × (1-(device noncovered time)/total study time).

We calculated compliance separately for each individual and each device using the millisecond coverage technique. This technique is designed to account for the slight variability in the rate at which measurements are taken by measuring the percentage of on-study time that is within T (defined above) seconds of a valid measurement. Compliance was estimated as the proportion (%) of on-study milliseconds within T seconds of a valid measurement.

Summary statistics

Computation of summary statistics across subjects and time points allowed us to explore the reasonableness of the data. We computed the arithmetic mean, SD, and minimum and maximum for all data sources and individuals. In addition, we computed a measurement timeline, averaged across individuals for the full study. This allowed us to explore any diurnal patterns in the data. All statistical analysis was done using the R software package version 3.3.2 with software libraries "plyr" (https://cran.r-project.org/web/packages/plyr/index.html) for data processing and "lme4" (https://cran.r-project.org/web/packages/lme4/index.html) for fitting of linear mixed models.

Assessment of diurnal variation

The degree to which measurements varied as a function of time of day was analyzed by calculating the minute-by-minute averages of HR, RR, and ST measurements and plotting these as a function of time of day. For each measurement, we calculated the minute of the day when the measurement was made. All measurements for the same minute were combined; for example, to calculate the "average HR" for 8:01 AM, all HR measurements taken between 8:01 AM and 8:02 AM were examined for all study days and for all subjects. Measurements were also grouped into "daytime" and "nighttime" periods. For this analysis, daytime was defined as between 8:00 AM and 9:00 PM and nighttime as between 12:00 AM and 6:00 AM. The periods between 6:00 AM and 8:00 AM and between 9:00 PM and 12:00 AM were anticipated to be "grey areas" with considerable variability within and between the subjects as to awake or asleep status during these periods.

Comparison between conventional in-clinic and wearable device measures

We compared the HR, RR, and ST measurements reported by the HealthPatch to the time-matched clinic measurements of HR, RR, and core body temperature (BT), respectively. In-clinic HR was collected using the Dinamap device. The study site's electronic source data system automatically captured the procedure timestamp at the time of collection. BT was collected using an electronic oral thermometer again linked to the Dinamap unit. In-clinic RR was collected manually by the site staff by observing the subjects' chest wall movements and counting respiration cycles over a defined period and entered immediately into the site's system, together with the time of data entry. We mapped the corresponding data from the wearable devices to the in-clinic data and then assessed the degree of concordance between the mapped data points at matched time points. The degree of concordance between in-clinic and wearable device data was determined using three separate strategies: correlation, regression, and Bland-Altman analyses. First, we calculated the Pearson correlation coefficient between the in-clinic and mapped wearable measurements. For regression, we performed ordinary least squares regression with the in-clinic measurement as the independent variable and the wearable measurement as the dependent variable.

For the Bland-Altman analyses,[14] we produced Bland-Altman Mean (BAM) and Bland-Altman Difference (BAD) values. For each in-clinic measurement, BAM was the average of the in-clinic and mapped wearable measurements, and BAD was the difference between the in-clinic and mapped wearable measurement. Bland-Altman plots were generated consisting of a scatterplot of BAM (x-axis) against BAD (y-axis). The points were color-coded by individual to help visualize any individual-specific bias. Computations of mean bias and 95% limits of agreement were also performed. The mean bias was simply the mean of BAD values. The 95% limits of agreement were calculated as 1.96 times the SD of BAD values.

Comparison between HealthPatch and Actiwatch actigraphy data

We investigated the extent of correlation between the HealthPatch step count and Actiwatch activity units. We divided each subject's time in the clinic into 1-hour intervals, beginning and ending on the hour. We then summed total activity units reported for each hour and calculated steps by subtracting step count at the start of the hour from step count at the end. Based on the epoch time (~1 second for HealthPatch and 30 seconds for Actiwatch), we calculated the number of measurements comprising complete data for 1 hour. We excluded any interval that was < 90% complete in either measurement and determined the Pearson correlation coefficient between activity units/hour and steps/hour.

RESULTS

Subject demographics

Six of the nine subjects enrolled in the core clinical study consented to participate in the exploratory wearable digital device evaluation component. Reasons given for nonconsent were the following: a history of prior severe cutaneous hypersensitivity to adhesives (one subject) and the perception that an honorarium payment should be offered by the sponsor for additional study procedures. The demographic

profile of the participants was as follows: five men and one woman; age range 18–55 years inclusive; three white subjects, two African-American subjects, one multiracial subject, and one smoker.

Completeness of data collection

For the HealthPatch, data completeness rates among the six subjects ranged from 83.6–99.2% (mean 93.1 ± 7.4%; **Table S1** and **Figure S2**). For the Actiwatch, completeness rates ranged from 62.6–98.6% (mean 88.9 ± 15%); the low rate (62.6%) occurred because the subject removed the device for the last 3 days of the study for unspecified reasons. Periods of loss of valid data for the HealthPatch device were attributed to poor skin contact and subjects removing the devices, again for unspecified reasons. Additional loss of valid Actiwatch data occurred due to device calibration issues and subjects removing devices without reporting this to the site personnel (**Figure S2**).

Comparison of in-clinic and wearable device measures

Comparison of the paired HR data showed a strong correlation between in-clinic and wearable measurements (Pearson's $r = 0.71$, $P = 2.2e\text{-}16$; **Figure 1**), confirmed by regression analysis ($\beta = 0.81$). Bland-Altman analysis (**Figure 1**) showed that, on average, in-clinic measurements were 0.91 bpm lower than their wearable counterparts. We estimated the 95% limits of agreement at 7.2 bpm, corresponding to 11% of the mean HR. Overall, the HR wearable-device data correlated well with the traditional in-clinic counterpart.

Recordings for RR derived from the HealthPatch seemed to be substantially different from the corresponding in-clinic measures. There was no significant relationship between the in-clinic and wearable device measurements (**Figure 1**) by either correlation (Pearson's $r = 0.08$, $P = 0.44$) or regression analysis ($\beta = 0.14$). Bland-Altman analysis corroborated these findings, with 95% limits of agreement of 6.0 breaths/min corresponding to 35% of the mean RR, indicating that the RR reported by the HealthPatch were statistically independent of the RR measured in-clinic.

There was no significant relationship between the ST as reported by the HealthPatch and in-clinic oral BT by either correlation (Pearson's $r = 0.14$, $P = 0.16$) or regression analyses ($\beta = 0.31$). Bland-Altman analysis indicated that ST and oral temperature had different distributions (**Figure 1**).

Actigraphic mobility and sleep data

The actigraphy data indicated much lower movement activity during the nighttime period, as expected. **Table 1**

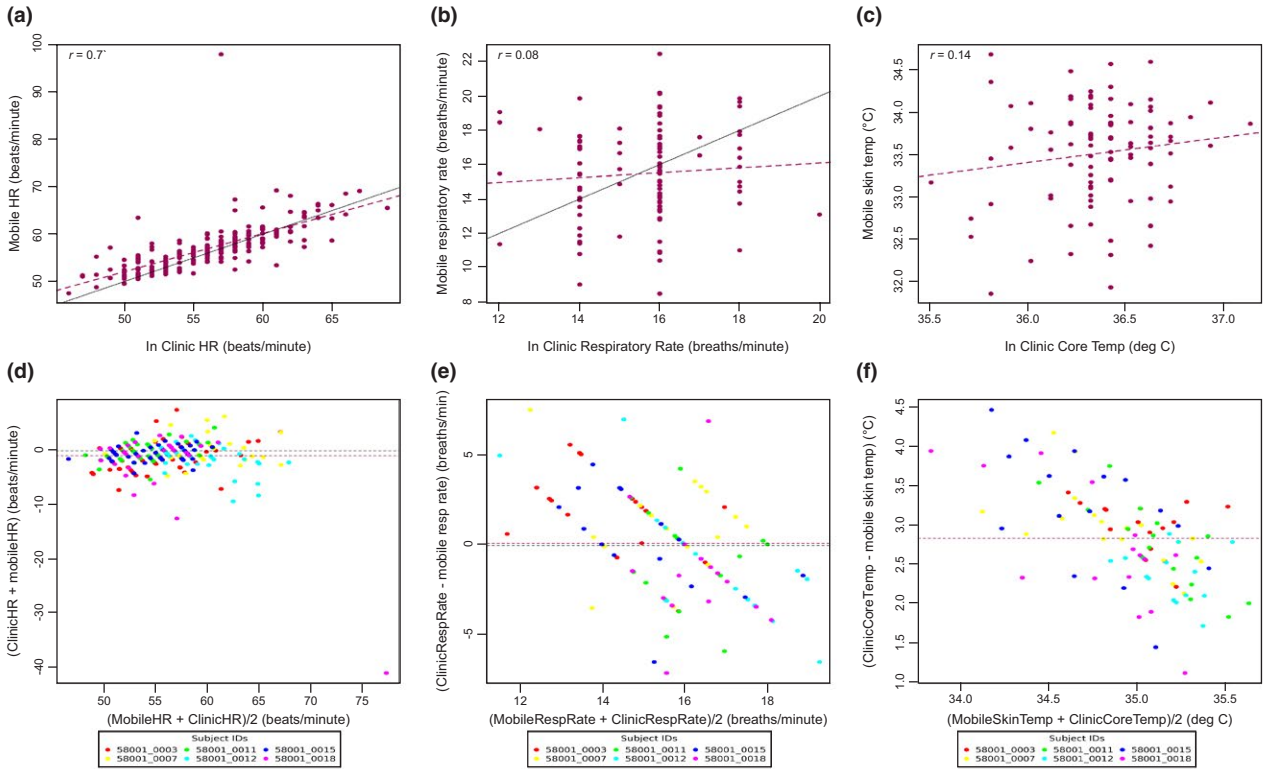

Figure 1 The comparison of in-clinic and wearable measurements for heart rate (HR), respiratory rate (RR), and skin temperature (ST) by correlation, regression (**a**–**c**), and Bland-Altman (**d**–**f**) analyses. The solid lines in **a**, **b**, and **c** are the line $y = x$, which would be expected if concordance were perfect; the dotted line is the actual regression line. In **c** there is no solid line because ST is significantly lower than core temperature. The Bland-Altman analysis results are presented in **d**, **e**, and **f**. The black dotted line is the line $y = 0$ indicating no mean difference between the two measures. The red dotted line is the mean difference line based on the actual data. **f** ST there is no black dotted line because ST is significantly lower than core temperature. **a** and **d** show the analyses for HR, **b** and **e** shows RR, **c** and **f** shows ST. Points are color-coded with a unique color for each subject.

Table 1 Summary statistics for activity measurements and total sleep time derived from the Actiwatch data

Subject ID	Daytime activity, counts/minute min–max	Nighttime activity, counts/minute min–max	TST, minutes min–max	TST, minutes mean ± SD
58001_0003	0–1132	0–801	196–501	367.9 ± 89
58001_0007	0–1453	0–514	190.5–536	376.5 ± 126.7
58001_0011	0–1499	0–584	306–471.5	367.5 ± 51.4
58001_0012	0–1063	0–908	118.5–383.5	268.6 ± 90.3
58001_0015	0–1697	0–998	313–498.5	409.8 ± 59.5
58001_0018	0–1697	0–1063	242–439	362.6 ± 68.7
OVERALL	0–1697	0–1063	118.5–536	358.8 ± 91.1

TST, total sleep time.
Range by endpoints for measurements of daytime activity, nighttime activity, and TST.

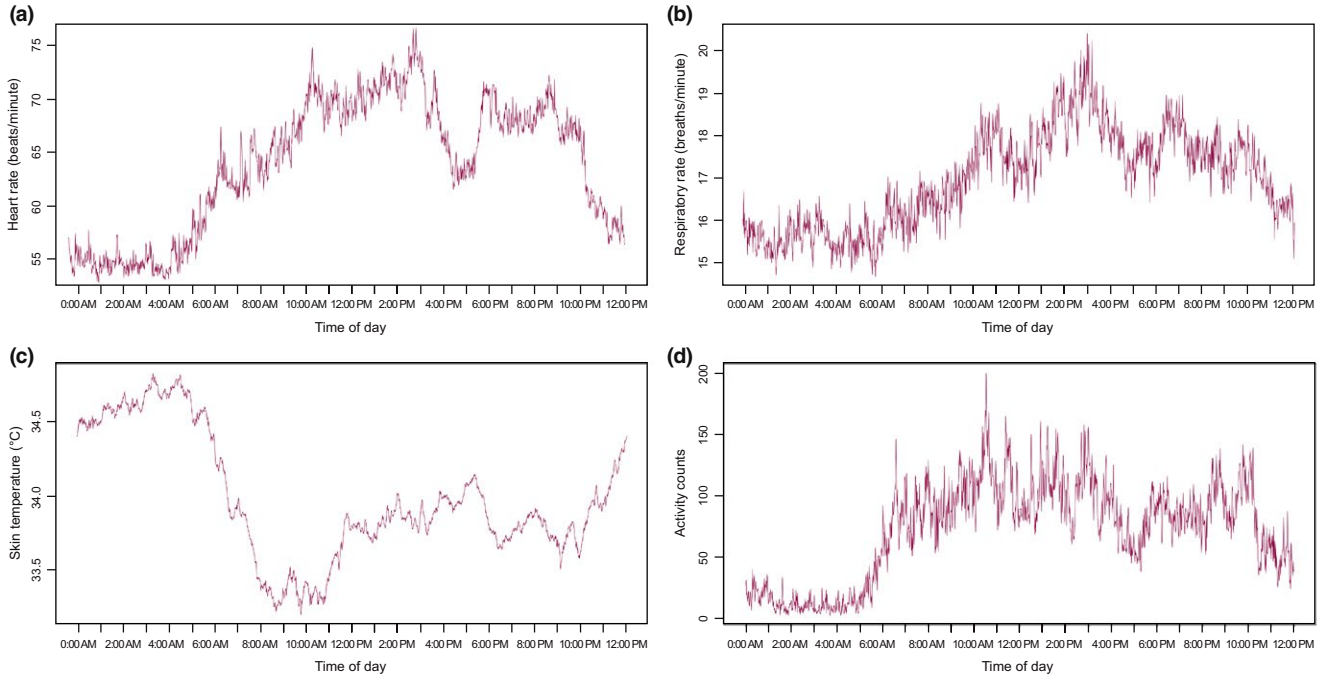

Figure 2 The aggregate diurnal patterns for heart rate (**a**), respiratory rate (**b**), skin temperature (**c**), and activity counts (**d**).

shows a summary of the data range (minimum and maximum) recorded by the Actiwatch. Analysis of total sleep time (**Table 1**) suggested that the subjects' mean sleep time was of 358 ± 91.1 minutes, or ~6 hours per night, with substantial variation between subjects.

Additionally, we compared actigraphy data derived from both HealthPatch (chest worn) and Actiwatch (wrist worn) devices by examining correlation of corresponding device outputs: step counts and activity counts. The comparison of these activity measures indicated that the readouts from these devices were broadly in agreement (**Figure S3**). The degree of correlation varied among study subjects (**Table S2**). However, we consistently observed some low values reported from Actiwatch devices corresponding to 0 values reported by the HealthPatch device (**Figure S3**) in all subjects.

Diurnal patterns

HR, RR, ST, and movement activity all demonstrated significant diurnal variation. HR, RR, and activity displayed very similar temporal patterns: lowest at night, highest in the afternoon, and an early evening nadir (**Figure 2**), as expected. The ST showed a different pattern; we observed the highest ST at night followed by a sharp drop in the morning.

Vital sign data

Most HR values reported by the HealthPatch were within or close to the normal range for healthy adults at rest, ~50–100 beats/min (bpm). Similarly, most reported RR recordings were within the expected physiological range (**Figure 3**). Most measurements for ST were within the range of previously reported for healthy adult normal ST (33–35°C; **Figure 3**) but were significantly different from

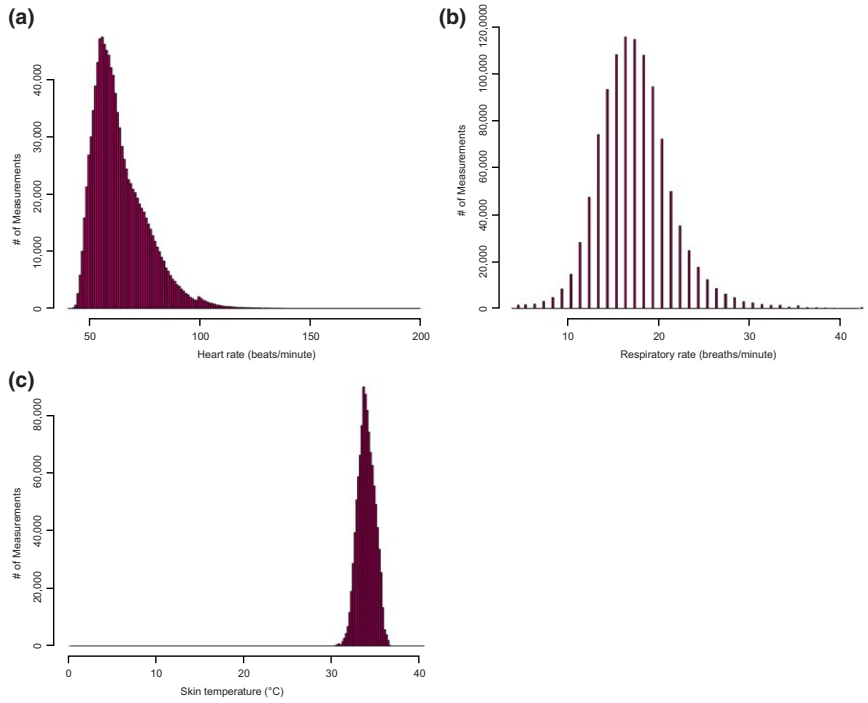

Figure 3 The histogram depicting the range of value distribution for heart rate (**a**), respiratory rate (**b**), and skin temperature (**c**).

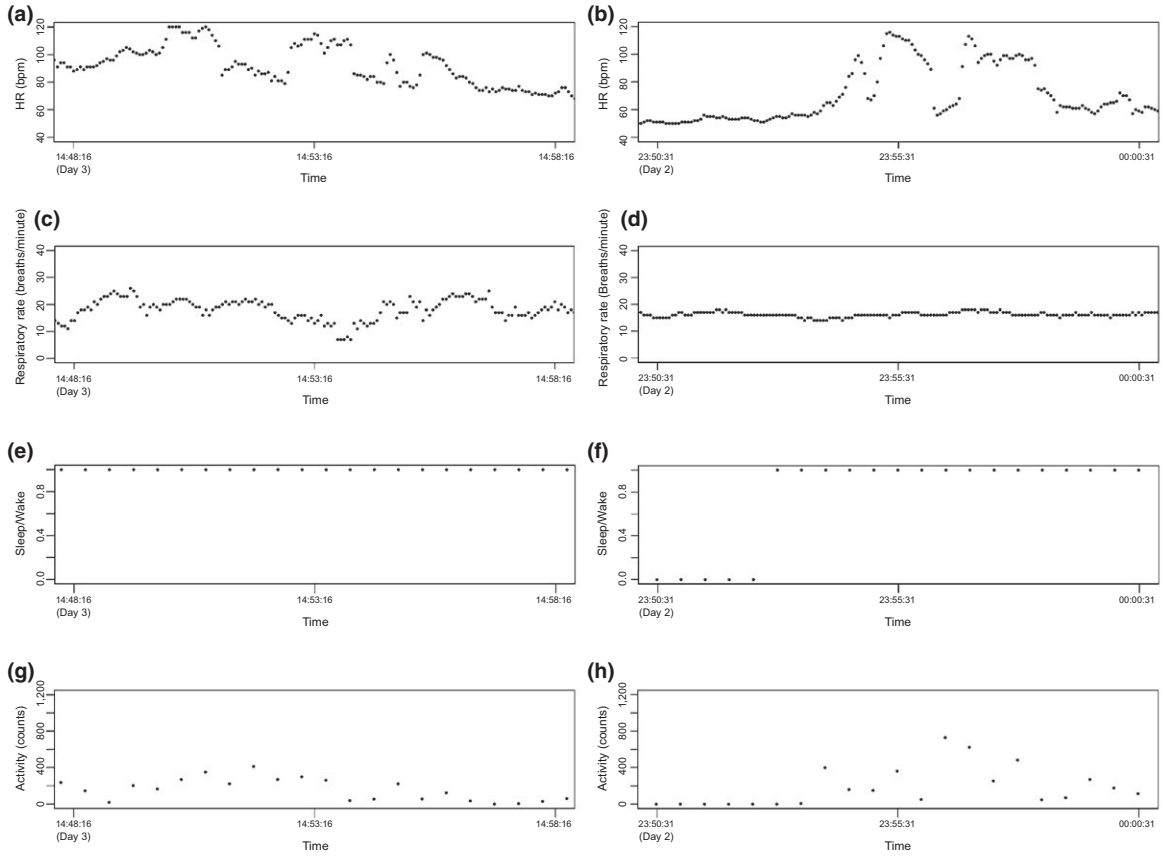

Figure 4 Visualization of heart rate (HR) (**a**, **b**) and respiratory rate (**c**, **d**) temporal patterns in the context of wake–sleep data (**e**, **f**), and subject mobility as assessed by activity counts (**g**, **h**) for subjects 58001_0003 (**a**, **c**, **e**, and **g**) and subject 58001_0011 (**b**, **d**, **f**, and **h**). bpm, beats/min.

the body core temperature as measured with the oral thermometer at the same time point.

Investigation of reported abnormal HR and RR values

To aid interpretation of HR and RR values that were outside of the expected resting range we compared the temporal change patterns of HR and RR as reported by the HealthPatch with activity level and wake-sleep status derived from the Actiwatch. Elevated readings were separated into two categories: (i) modestly elevated values that might be readily ascribed to changes in body position, physical activity, and/or study procedures, and (ii) significantly elevated values – those that would require further attention as they could indicate potential drug-related safety signals should they occur in a phase I setting and be temporally related to exposure to the study drug or to a challenge agent. Readouts for the same period were examined graphically (**Figure 4**), revealing that most episodes of elevated HR and RR occurred at times of increased physical activity. We found that comparing time-matched data by direct visual comparison of HR and RR data on the one hand and activity counts and wake-sleep status on the other was informative for interpreting moderately abnormal readings. However, this visual review process was time-consuming.

There were several episodes when the HR was elevated above 150 bpm or even 180 bpm, which exceeds reported "normal" HR ranges for a comparable healthy volunteer study population.[15] This finding was unexpected given that the study recruited healthy volunteer subjects, that the study subjects were confined to the CPU for the duration of the wearable device evaluation, that no cardiac-related adverse events were expected based on the properties of the investigational compound or were detected during the study by conventional means, and that, in compliance with the study protocol, the subjects were restricted from strenuous physical exercise. Manual reviews of the ECG waveforms were done for periods during which HR values were below 50 bpm or above 180 bpm. This indicated that all device-reported values below the normal range were consistent with the reported in-clinic HR values (**Figure 5a**), consistent with sinus bradycardia occurring in healthy individuals.

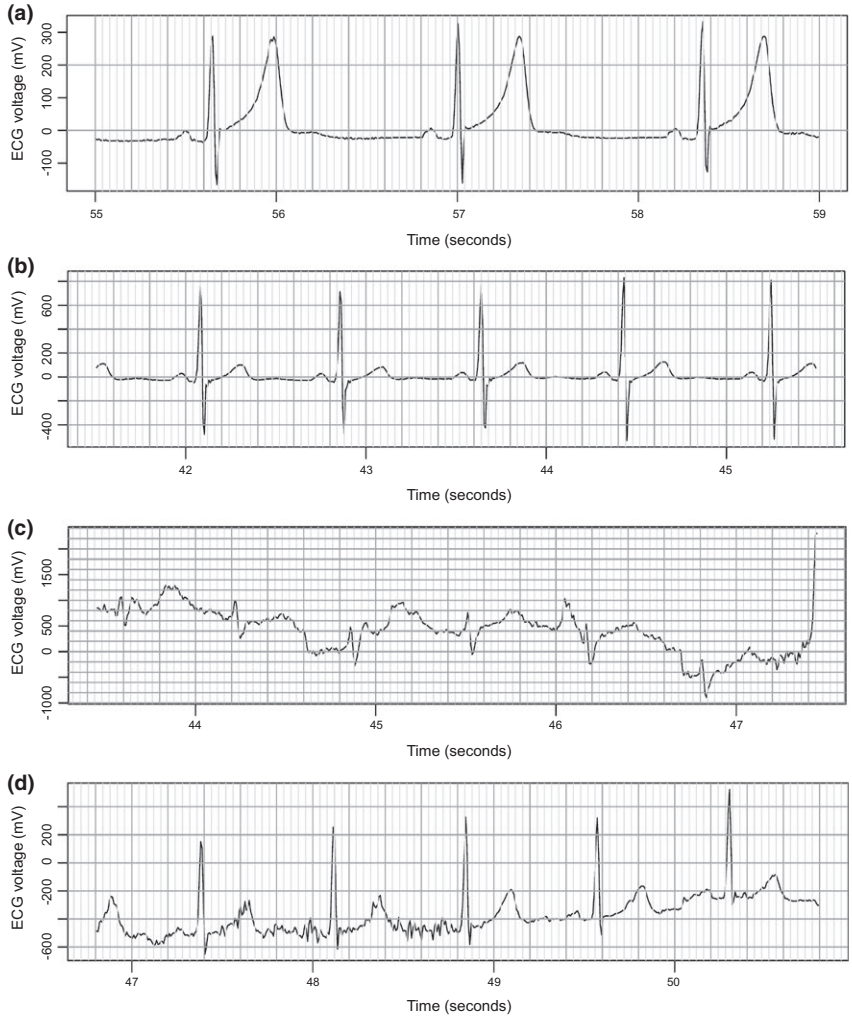

Figure 5 Visual examination of the echocardiogram (ECG) waveforms for episodes with reported heart rates of 44 bpm (**a**), 79 bpm (**b**), 203 bpm (**c**), and 193 bpm (**d**), bpm, beats/min.

However, examination of single-lead ECG strips corresponding to reported HR values above 180 bpm revealed artefacts of data recording and data processing (**Figure 5, c and d**) rather than true episodes of tachycardia. This additional investigation of these reported periods of out-of-range HR was resource-intensive, requiring manual visual review by a physician of each ECG tracing corresponding to these periods.

The number of time intervals containing values outside of normal range that would require follow-up was determined by calculating the number epochs with HR values above 150 bpm and above 180 bpm (**Table 2**). The number of such epochs was highly variable between subjects, with the highest number for subject 58001_0018. We also calculated the number and total duration of gaps in data collection in order to estimate the overall amount of vital sign data not being collected, thus providing an estimate of the likelihood of missing a safety signal if the device had been used as the primary method for collecting vital sign data. **Table S3** shows the number of time intervals and a total duration of such intervals, with the highest number occurring in subject 58001_0018. We found significant gaps in the completeness of vital sign data collection, with between 0.6 and 35.4 hours of data missing over the course of the 10-day study period.

The reported minimum ST values of 0°C were likely due to poor patch adherence or other artefact, despite the apparently valid skin impedance measures.

Study subject and site personnel feedback

At the end of the study, both the study participants and the study coordinators were asked to complete a brief satisfaction questionnaire to assess their comfort level with the device technology used and their willingness to participate in similar study procedures in the future. The majority of the subjects indicated a high level of acceptability for the devices and a willingness to participate in future studies that assessed wearable digital device technologies.

Overall, study coordinators at the CPU reported high levels of satisfaction on the training and technical support they receive. However, in free-form feedback the study coordinators did highlight a desire to gain additional "hands-on" experience with devices and their associated software in order to increase their comfort level.

DISCUSSION

This study assessed the feasibility of using 510(k)-cleared wearable digital devices for collecting physiological and activity data in the context of a residential drug development clinical study together with the performance and fit-for-purpose validation of these devices. The HealthPatch and Actiwatch devices used did not interfere with the other study procedures, such as dosing, safety, and pharmacokinetic sample collection and were well received by the study subjects and the site personnel. Activity counts and sleep duration data derived from the Actiwatch device had face validity, followed expected diurnal patterns, and were consistent with previously published results.[16] The average total sleep time of 6 hours is perhaps shorter than a typical night's sleep for a healthy adult; this may be a consequence of the unfamiliar environment of the CPU. The observed diurnal patterns in HR, RR, ST, and activity were consistent with the previous reports.[17,18] Our experience of periods of missing vital sign data is consistent with previously published results.[8]

Table 2 Quantification of number and percent of epochs with HR values above 150 and 180 bpm

Subject	Number (%) of epochs with HR > 150 bpm	Number (%) of epochs with HR > 180 bpm
58001_0003	29 (0.02)	13 (0.01)
58001_0007	18 (0.01)	5 (0.00)
58001_0011	9 (0.00)	6 (0.00)
58001_0012	67 (0.04)	6 (0.00)
58001_0015	99 (0.06)	33 (0.02)
58001_0018	258 (0.16)	77 (0.05)

bpm, beats/min; HR, heart rate.

There were significant limitations with the HR data produced by the HealthPatch device because of the volume of artefacts produced that require a follow-up and a manual review. Although the HR recordings showed good correlation with traditional in-clinic measures, there were many reported episodes of tachycardia due to voltage artefacts that required manual visual review by a physician to resolve.

Differences between RRs as determined by the device and by manual in-clinic measures may in fact reflect the differences between manual and device-mediated methods, which have been described by other groups.[19] Differences between surface ST and body core temperature are expected.[20] ST is typically lower and more variable than BT[20,21] and is affected by the site of measurement, clothing, environmental temperature, and even emotional state. However, we expected some degree of relationship between these two variables. In general, we observed much smaller variation in core temperature than ST, as expected, which further underscores the distinction between the two parameters.

The comparison of activity data generated by HealthPatch and Actiwatch devices revealed that the measurements were broadly in agreement although not perfectly correlated. Less than perfect correlation was expected because the devices are located on different parts of the body (trunk vs. wrist); in addition, different activity readouts (step counts vs. activity counts) impacted the types of physical activities detected (i.e., walking vs. moving the upper body only). Given that the Actiwatch readout provided data for a wider variety of physical activities, the activity counts derived from Actiwatch were used for vital sign data interpretation. These differences illustrate the need for data standardization for similar device readouts (e.g., variables associated with actigraphy).

The extent to which each device was evaluated was driven by the intended use of the data. We applied more rigor for the HealthPatch data analysis because of its potential to detect a safety signal. The data derived from the Actiwatch device played a secondary role and were largely used to interpret the vital sign values outside of the normal range.

The device evaluation portion of this study has several limitations. In the broad context of drug development, the phase I study in which these devices were evaluated was of low complexity in terms of device implementation logistics,

data analysis, and interpretation. The study subjects were inpatients and were monitored by the CPU personnel for the entire duration of the study, which facilitated data interpretation and helped to qualify certain findings as artefacts. In addition, the small size of the study ($N = 6$) allowed the many reports of out-of-range vital signs (e.g., potential tachycardia episodes) to be reviewed manually. This may prove to be too resource-intensive in a larger study. Our findings demonstrated significant challenges with continued device use, data collection, processing, and, most importantly, data interpretation. We anticipate that the impact of these issues would be even higher with a study involving more subjects, if additional procedures were included (e.g., imaging or invasive sampling) or if the device component of the study was done remotely with study subjects wearing the devices at home. We also observed missing data, an effect that is likely to be amplified in study subjects with medical conditions who are seen in usual practice. Adherence was an issue in this study of relatively low complexity, although it was comparable with other similar reports[8]; it is likely to be a limiting factor in subjects with disease conditions as well. In addition, our ability to review information was somewhat restricted, as the device manufacturer uses proprietary software algorithms for initial data processing.

Further limitations on the use of wearable devices in drug-development studies are that with the model of the device evaluated in this study and the available analysis software packages: (i) the derivative data were not available in "real time" during the study, which would delay the detection of acute safety signals and would not enable the investigator or sponsor to make real-time clinical decisions on the management of safety issues, and (ii) the single-lead ECG information that is reported is limited to HR, and, in particular, does not provide interpretation of potential rhythm abnormalities or of important electrocardiographic intervals, such as the QTc time. However, for arrhythmia detection, a single-lead ECG device has the advantages of ease of use and convenience (compared with conventional Holter monitoring[22]) and in this context of use, might be fit-for-purpose with appropriate software development. The use of a chest wall patch device to detect atrial fibrillation in a pragmatic population-based study of over 2,000 subjects was recently reported.[23] In addition, our study demonstrated the critical need for access to the source data in order to evaluate study results, to confirm reports of abnormal activity, and to understand the data limitations. Source data and algorithm transparency remain an issue with both consumer grade and some medical grade devices in the context of clinical investigations.

Our findings indicate the need for careful evaluation of wearable digital devices according to the fit-for-purpose principle before the device-derived data can be used to support primary or secondary study endpoints, irrespective of the regulatory clearance. Regulatory evaluation of device performance under the auspices of the FDA 510(k) clearance program conveys a level of reassurance regarding the technical performance of a device. However, receipt of 510(k) clearance should not be taken to imply that the device is fit-for-purpose for an industry-sponsored drug development study. We did not clinically validate the HealthPatch as fit-for-purpose for augmented physiological data collection because of artefacts, including false-positive HR signals and missing data. This issue of false-positive signals is not inherent to any specific device. Several groups have previously reported a false-positive rate from ECG monitors in the intensive care unit setting as high as 75–93%.[24] Our finding of poor correlation between device-reported and manual in-clinic measurements of RR is also consistent with results reported previously.[19] Nonetheless, the issue of specificity of safety monitoring limits the potential utility digital devices for drug development. We believe that a device similar to the HealthPatch device could be of utility for monitoring study subjects if the issues of false-positive results and missing data are addressed to an acceptable extent. The data derived from such a device can be used in a manner similar to an "early warning score" system[25] to generate signals to be investigated further and facilitate building an investigational drug safety profile early during clinical development.

There are many promising uses of wearable devices in clinical trials as well as several challenges. Potential applications drive toward an enhanced understanding of disease variability, treatment response, safety assessment, innovation in clinical trial design and conduct, as well as increasing efficiency and decreasing costs in clinical trials. Although the promises are clear, the challenges are not insignificant and include scientific, regulatory, ethical, legal, data management, infrastructure, analysis, and security challenges.[5] The current study demonstrates practical examples of scientific/regulatory, data management, infrastructure, and analysis issues, as described above. We did not encounter significant obstacles with ethical, legal, or security issues, but the importance of these may have been diminished by the pilot nature of this substudy.

In summary, comparison between specific wearable digital devices and in-clinic measures established a strong correlation for HR but poor correlation between in-clinic and wearable measurements of RR and of ST using the HealthPatch. We concluded that the HealthPatch was not fit-for-purpose for HR monitoring because of the artefacts it produced and the amount of time required for data processing and review. The number of artefacts would need to be greatly reduced before a wider of implementation of this device in clinical trials. The Actiwatch device was used as a supporting application to interpret the vital sign data and was suitable for the intended purpose of monitoring movement, aiding interpretation of abnormal vital sign data, and collecting certain sleep parameters. For wearable devices to gain wider applicability in drug development, we need to develop and establish acceptance for common issues, including medical need, device choice, context of use, fit-for-purpose validation, and predefined operational requirements, as well as data collection, processing, and interpretation. Careful consideration must be given to clinical validation and context of use to assure that device measurements are fit-for-purpose. Clinical Trial Transformation Initiative made substantial progress in addressing these issues by developing recommendations for implementation of mobile technologies in human experimentation.[26] However, there is a great need to supplement these recommendations with the results derived from clinical studies. The current study illustrates

the critical role for evaluation, both analytical and clinical, in the applicability of wearable devices.

Supporting Information. Supplementary information accompanies this paper on the *Clinical and Translational Science* website (www.cts-journal.com).

Figure S1. Data flow diagrams for the Actiwatch and the HealthPatch devices in the phase I study.
Figure S2. Completeness of data collection for the HealthPatch (**a**) and Actiwatch (**b**) devices.
Figure S3. The comparison of activity measurements by the HealthPatch (step counts, x-axis) and Actiwatch (activity counts, y-axis) devices for study subjects 58001_003 (**a**), 58001_007 (**b**) 58001_0011 (**c**), 58001_0012 (**d**), 58001_0015 (**e**), and 58001_0018 (**f**).
Table S1. Completeness of data collection for the HealthPatch and Actiwatch devices for each study subjects.
Table S2. Correlation of activity measurements by the HealthPatch and Actiwatch devices.
Table S3. Quantification of time intervals with missing data.

Acknowledgments. The authors thank Kristina Allikmets and Jason Homsy for help with the data review and interpretation.

Funding. No funding was received for this work.

Conflict of Interest. E.S.I., I.L.M., G.H., D.M., E.D.P., and J.A.W. are employees of Takeda Pharmaceuticals and own stock in the company. G.B., M.C., and C.B. are employees of Koneksa Health and own stock in the company. As Editor-in-Chief for *Clinical and Translational Science*, J.A.W. was not involved in the review or decision process for this paper.

Author Contributions. E.S.I., I.L.M., G.B., D.M., and J.A.W. wrote the manuscript. E.S.I., I.L.M., G.B., M.C., E.D.P., C.B., and J.A.W. designed the research. E.S.I., I.L.M., G.B., G.H., M.C., and C.B. performed the research. E.S.I., I.L.M., G.B., G.H., and D.M. analyzed the data.

1. Mealer, M. *et al.* Remote source document verification in two national clinical trials networks: a pilot study. *PLoS One* **8**, e81890 (2013).
2. Wu, T.C. *et al.* Telemedicine-guided remote enrollment of patients into an acute stroke trial. *Ann. Clin. Transl. Neurol.* **2**, 38–42 (2015).
3. Smalley, E. Clinical trials go virtual, big pharma dives in. *Nat. Biotechnol.* **36**, 561–562 (2018).
4. Rosenbaum, L. Swallowing a spy - the potential uses of digital adherence monitoring. *N. Engl. J. Med.* **378**, 101–103 (2018).
5. Izmailova, E.S., Wagner, J.A. & Perakslis, E.D. Wearable devices in clinical trials: hype and hypothesis. *Clin. Pharmacol. Ther.* **104**, 42–52 (2018).
6. Clifton, L., Clifton, D.A., Pimentel, M.A., Watkinson, P.J. & Tarassenko, L. Predictive monitoring of mobile patients by combining clinical observations with data from wearable sensors. *IEEE J. Biomed. Health Inform.* **18**, 722–730 (2014).
7. Roess, A. The promise, growth, and reality of mobile health - another data-free zone. *N. Engl. J. Med.* **377**, 2010–2011 (2017).
8. Weenk, M. *et al.* Continuous monitoring of vital signs using wearable devices on the general ward: pilot study. *JMIR Mhealth Uhealth.* **5**, e91 (2017).
9. Schmier, J.K. & Halpern, M.T. Patient recall and recall bias of health state and health status. *Expert Rev. Pharmacoecon. Outcomes Res.* **4**, 159–163 (2004).
10. Lee, J.W. *et al.* Fit-for-purpose method development and validation for successful biomarker measurement. *Pharm. Res.* **23**, 312–328 (2006).
11. National Center for Biotechnology Information. BEST (Biomarkers, EndpointS, and other Tools) Resource <https://www.ncbi.nlm.nih.gov/books/NBK326791/>. (January 2016). Accessed November 30, 2018.
12. US Food and Drug Administration. 510(k) summary. <https://www.accessdata.fda.gov/cdrh_docs/pdf/K983533.pdf>. (January 2014). Accessed November 30, 2018.
13. US Food and Drug Administration. 510(k) summary. <https://www.accessdata.fda.gov/cdrh_docs/pdf13/K132447.pdf>. (April 2014). Accessed November 30, 2018.
14. Bland, J.M. & Altman, D.G. Measuring agreement in method comparison studies. *Stat. Methods Med. Res.* **8**, 135–160 (1999).
15. Sarzynski, M.A. *et al.* Measured maximal heart rates compared to commonly used age-based prediction equations in the heritage family study. *Am. J. Hum. Biol.* **25**, 695–701 (2013).
16. Thurman, S.M. *et al.* Individual differences in compliance and agreement for sleep logs and wrist actigraphy: a longitudinal study of naturalistic sleep in healthy adults. *PLoS One* **13**, e0191883 (2018).
17. Bracci, M. *et al.* Peripheral skin temperature and circadian biological clock in shift nurses after a day off. *Int. J. Mol. Sci.* **17**, (2016).
18. Degaute, J.P., van de Borne, P., Linkowski, P. & Van Cauter, E. Quantitative analysis of the 24-hour blood pressure and heart rate patterns in young men. *Hypertension* **18**, 199–210 (1991).
19. Smith, I., MacKay, J., Fahrid, N. & Krucheck, D. Respiratory rate measurement: a comparison of methods. *Br. J. Healthcare Assistants* **5**, 18–22 (2011).
20. Xu, X., Karis, A.J., Buller, M.J. & Santee, W.R. Relationship between core temperature, skin temperature, and heat flux during exercise in heat. *Eur. J. Appl. Physiol.* **113**, 2381–2389 (2013).
21. Choi, J.K., Miki, K., Sagawa, S. & Shiraki, K. Evaluation of mean skin temperature formulas by infrared thermography. *Int. J. Biometeorol.* **41**, 68–75 (1997).
22. Barrett, P.M. *et al.* Comparison of 24-hour Holter monitoring with 14-day novel adhesive patch electrocardiographic monitoring. *Am. J. Med.* **127**, 95e11–95e97 (2014).
23. Steinhubl, S.R. *et al.* Effect of a home-based wearable continuous ECG monitoring patch on detection of undiagnosed atrial fibrillation: the mSToPS randomized clinical trial. *JAMA* **320**, 146–155 (2018).
24. Tsien, C.L. & Fackler, J.C. Poor prognosis for existing monitors in the intensive care unit. *Crit. Care Med.* **25**, 614–619 (1997).
25. Mitchell, I.A. *et al.* A prospective controlled trial of the effect of a multi-faceted intervention on early recognition and intervention in deteriorating hospital patients. *Resuscitation* **81**, 658–666 (2010).
26. Clinical Trial Transformation Initiative (CTTI). Program: Mobile Clinical Trials (MCT). <https://www.ctti-clinicaltrials.org/projects/mobile-technologies>. (July 2018). Accessed November 30, 2018.